

LOW VOLTAGE NON-VOLATILE MEMORY TRANSISTOR

Kevin T. Look

FIELD OF THE INVENTION

[0001] The present invention relates to non-volatile memory transistors. More specifically, the present invention relates to a low-voltage non-volatile memory transistor that can be fabricated using a standard CMOS process.

RELATED ART

[0002] As semiconductor process technology advances, both vertical and horizontal dimensions of the resulting devices are scaled to smaller and smaller dimensions. The operating voltage of the resulting field effect transistors is also reduced to avoid a high electric field across the transistor gate oxide. However, the voltage required to turn the transistor on or off (hereinafter referred to as the threshold voltage), has not been scaled down as aggressively as the operating voltage. The difference between the operating voltage and the threshold voltage becomes smaller and smaller as semiconductor process technology progresses. In today's deep sub-micron technology (e.g., 90 nanometers (nm) or less), the operating voltage of the resulting device is reduced to about 1.0 Volt while the threshold voltage of the associated transistors remains at about 0.4 Volts. (i.e., the same threshold voltage associated with process technologies of 130 nm or greater). Certain operating conditions may undesirably shift the threshold voltage of the transistors close to the operating voltage, such that it is difficult to turn on the transistor.

[0003] The threshold voltages of P-channel transistors are known to degrade when exposed to a negative bias voltage at a

high temperature. The temperature effect on a negative biased p-channel transistor is described by Yamamoto et al. in "Bias Temperature Instability in Scaled p+ Polysilicon Gate p-MOSFET's", IEEE Transactions on Electron Devices, Vol. 46, No. 5, May 1999, pp. 921-926. Yamamoto et al. indicates that the threshold voltage shift is an exponential function of the inverse of the temperature. Extrapolation from the data presented by Yamamoto et al. shows that the threshold voltage can shift by more than 1 Volt at temperatures over 500°C. Yamamoto et al. also indicates that the threshold voltage shift is dependent on the channel length of the transistor.

[0004] In "NBTI Enhancement by Nitrogen Incorporation into Ultrathin Gate Oxide for 0.10-um Gate CMOS Generation", 2000 Symposium on VLSI Technology Digest of Technical Papers, p. 93, Kimizuka et al. also reported that the lifetime of a p-channel transistor (i.e., change in threshold voltage) is an exponential function of the inverse of the temperature.

[0005] Titanium (Ti) polysilicide and cobalt (Co) polysilicide are commonly used materials for transistor gates and interconnects in integrated circuit manufacturing processes due to the low resistivity exhibited by these materials. The effect of the required anneal temperature on the resistivity of titanium silicide has been described by Lasky et al. in "Comparison of Transformation to Low-Resistivity Phase and Agglomeration of TiSi_2 and CoSi_2 ", IEEE Transactions on Electron Devices, Vol. 38, No. 2, February 1991, pp. 262-269. Lasky et al. teaches that the resistivity of titanium silicide is transformed from a low resistivity (5-15 Ohms/sq.) to a high resistivity (greater than 100 Ohms/sq.) at an anneal temperature greater than 950°C. Lasky et al. also attributes the agglomeration of the titanium silicide to the change in resistivity at these high temperatures.

[0006] The effect of the thermal process on the performance of a titanium silicide gate device has been described by Jang et al. in "Effects of Thermal Processes After Silicidation on the Performance of TiSi_2 /Polysilicon Gate Device", IEEE Transactions on Electron Devices, Vol. 46, No. 12, Dec. 1999, pp. 2353-2356. Jang et al. observes that the resistivity of a titanium silicide gate changes as the BPSG reflow temperature increases from 750°C to 850°C.

[0007] In "A PROM Element Based on Salicide Agglomeration of Poly Fuses in a CMOS Logic Process", International Electron Devices Meeting (IEDM) Technical Digest, pp. 855-858, 1997, Alavi et al. describes a structure that can be programmed to change resistivity by passing current through a titanium silicide element in the structure, thereby causing agglomeration. The agglomeration is initiated with a low voltage (less than 2 Volts) and a moderate current (less than 10 mA). The change in resistivity is an indication that passing current through the titanium silicide structure generates enough heat to reach the critical temperature to start the agglomeration.

[0008] Similarly, a resistivity change in a cobalt silicide fuse structure has been demonstrated by Kalnitsky et al. in "CoSi₂ Integrated Fuses on Poly Silicon for Low Voltage 0.18 μm CMOS Applications", International Electron Devices Meeting (IEDM) Technical Digest, pp. 765-768, 1999.

[0009] It would therefore be desirable to have a structure that takes advantage of (1) the P-channel transistor's threshold voltage instability at negative bias and high temperature and (2) the high temperature generation in the presence of current.

SUMMARY

[0010] Accordingly, the present invention provides a 5-terminal p-channel non-volatile memory (NVM) transistor that is programmed by shifting the threshold voltage of the transistor. The threshold voltage is shifted by introducing a programming current to the gate electrode of the transistor, and simultaneously introducing a negative bias to the transistor.

[0011] In accordance with one embodiment, the NVM transistor includes an n-type semiconductor body region, first and second p-type source/drain regions located in the semiconductor body region and separated by a channel region, and a polysilicon/metal polysilicide gate electrode having a central portion located over the channel region, a first end portion located on a one side of the central portion, and a second end portion located on an opposite side of the central portion. First and second contact elements are coupled to the first and second source/drain regions, respectively. A third contact element is coupled to the first end of the gate electrode, and a fourth contact element is coupled to the second end of the gate electrode. A fifth contact element is coupled to the semiconductor body region.

[0012] The p-channel NVM transistor is programmed as follows. A first program control voltage, typically a core supply voltage (V_{CC}), or an input/output supply voltage (V_{DD}) is applied to the first end of the gate electrode via the third contact element. A second program control voltage, typically a ground supply voltage, is applied to the second end of the gate electrode via the fourth contact element. As a result, a programming current flows through the gate electrode, thereby generating heat within the gate electrode.

[0013] A third program control voltage, typically an input/output supply voltage (V_{DD}), is applied to the body region of the transistor. The third program control voltage can be the same or greater than the first program control voltage. As a result, the p-channel NVM transistor is partially or fully negatively biased.

[0014] The threshold voltage of the p-channel NVM transistor is shifted in response to the locally generated heat and the negative bias condition. The agglomeration of the metal polysilicide in the gate electrode is an indication of the intense heat generated during programming. As a result, the programming operation is permanent.

[0015] The p-channel NVM transistor is read by applying a first read control voltage, typically the ground supply voltage, to the first and second ends of the gate electrode, and to the first source/drain region of the transistor. A second read control voltage, typically the core supply voltage (V_{CC}), is applied to the semiconductor body region and the second source/drain region of the transistor. If the NVM transistor has been programmed, no read current flows between the first and second source/drain regions. Conversely, if the NVM transistor has not been programmed, a significant read current flows between the first and second source/drain regions.

[0016] The present invention will be more fully understood in view of the following description and drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] Fig. 1 is a layout diagram illustrating a five-terminal p-channel non-volatile memory transistor in accordance with one embodiment of the present invention.

[0018] Fig. 2A is a cross sectional view of the p-channel non-volatile memory transistor along section line A-A of Fig. 1.

[0019] Fig. 2B is a cross sectional view of the p-channel non-volatile memory transistor along section line B-B of Fig. 1.

DETAILED DESCRIPTION

[0020] The present invention provides a non-volatile memory (NVM) transistor that takes advantage of the threshold instability of a negatively biased p-channel transistor at a high temperature.

[0021] Fig. 1 is a layout diagram illustrating a five-terminal p-channel NVM transistor 100 in accordance with one embodiment of the present invention. Fig. 2A is a cross sectional view of p-channel NVM transistor 100 along section line A-A of Fig. 1. Fig. 2B is a cross sectional view of p-channel NVM transistor 100 along section line B-B of Fig. 1. As described in more detail below, p-channel NVM transistor 100 is controlled to operate as a non-volatile memory cell.

[0022] P-channel NVM transistor 100, which is fabricated in p-type substrate 201 and n-type well region 202, includes p-type source/drain regions 101-102, n-type well contact region 103, field dielectric region 105, gate electrode 110, electrically conductive contact elements 121-125, electrically conductive traces 131-135 and dielectric sidewall spacers 140. P-channel transistor 100 is fabricated using conventional CMOS processing techniques.

[0023] In the described embodiments, p-type substrate 201 is a monocrystalline semiconductor substrate (e.g., silicon). N-type well region 202 is formed in substrate 201 using conventional CMOS processing techniques (e.g., n-type ion

implantation). Field dielectric region 105 is formed in the upper surface of n-type well region 202 as illustrated. In the illustrated embodiments, field dielectric region 105 is a shallow trench isolation (STI) region, which is created by forming a trench in substrate 201, and then filling the trench with a dielectric material, such as silicon oxide. However, in other embodiments, field dielectric region 105 can be formed by the local oxidation of silicon (LOCOS).

[0024] A gate dielectric layer 106 is formed over the upper surface of n-type well region 202. The gate dielectric layer 106 can be, for example, thermally grown silicon oxide having a thickness in the range of about 10 to 25 Angstroms. A conductively doped polycrystalline silicon (polysilicon) layer 111 is formed over the upper surface of the gate dielectric layer 106. In the described embodiment, polysilicon layer 111 has a thickness in the range of about 500 to 1500 Angstroms. In this embodiment polysilicon layer 111 is doped with boron to a dopant concentration in the range of about 10^{20} atoms/cm³.

[0025] A layer of refractive metal polysilicide 112 is formed over the upper surface of polysilicon layer 111. In one embodiment, metal polysilicide layer 112 is formed by depositing a layer of refractive metal, such as titanium (Ti), nickel (Ni) or cobalt (Co), on polysilicon layer 111. A subsequent thermal anneal causes the refractive metal to react with the underlying polysilicon layer, thereby forming metal polysilicide layer 112. Alternately, a layer of metal polysilicide (TiSi or CoSi) can be deposited directly on polysilicon layer 111. In accordance with one embodiment, metal polysilicide layer 112 has a thickness in the range of about 1000 to 2000 Angstroms.

[0026] The metal polysilicide layer 112 and polysilicon layer 111 are patterned to form gate electrode 110. Gate

electrode 110 includes a first end 110A dimensioned to receive a first contact element 123, a second end 110B dimensioned to receive a second contact element 124, and a central region 110C dimensioned to define a channel region 104 of p-channel NVM transistor 100. Lightly doped source/drain regions (P-) are formed in n-well region 202, wherein the edges of these P- regions are self-aligned with edges of gate electrode 110. Dielectric sidewall spacers 140 (e.g., silicon nitride, silicon oxide, or silicon oxynitride) are formed adjacent to gate electrode 110. Heavily doped source/drain contact regions (P+) are formed in n-well region 202, wherein the edges of these P+ regions are self-aligned with edges of sidewall spacers 140. The P+ and P- regions form p-type source/drain regions 101 and 102. In an alternate embodiment, lightly doped P- regions and sidewall spacers 140 can be eliminated. Heavily doped n-well contact region 103 (N+) is formed in n-well region 202, away from source/drain regions 101-102. In one embodiment, self-aligned metal polysilicide (i.e., salicide) can be formed over exposed upper surfaces of source/drain regions 101-102, n+ type contact region 103 and polysilicon layer 111, in a manner known to those of ordinary skill in the semiconductor processing art.

[0027] Electrically conductive contact elements 121, 122, 123, 124 and 125 are formed through a first dielectric layer (not shown), thereby contacting source/drain region 101, source/drain region 102, the first end 110A of gate electrode 110, the second end 110B of gate electrode 110, and n+ contact region 103, respectively. Contact element 123 (coupled to first end 110A) and contact element 124 (coupled to second end 110B) may be coupled to separate nodes in a circuit, thereby allowing for independent control of first end 110A and second end 110B of gate electrode 110. In the described embodiment,

contact elements 121-125 are tungsten. However, other materials can be used to form contact elements 121-125 in other embodiments.

[0028] Electrically conductive traces 131-135, which are formed over the first dielectric layer, contact the contact elements 121-125, respectively. Although traces 131-135 are all illustrated as part of a first conductive layer (e.g., the first metal layer), it is understood that these traces may be located in different conductive layers in other embodiments, as required by the layout of the associated device.

[0029] In accordance with one embodiment, five-terminal p-channel NVM transistor 100 is controlled to operate as a non-volatile memory cell in the manner described below.

[0030] NVM transistor 100 is programmed by applying a first program control voltage to the first end 110A of gate electrode, a second program control voltage to the second end 110B of gate electrode 110, and a third program control voltage to the body/substrate (i.e., n-well 202) of NVM transistor 100. The first program control voltage is selected to be greater than the second program control voltage, such that a programming current flows through gate electrode 110, thereby generating localized heat. The third programming control voltage is selected to be greater than the second programming control voltage, such that p-channel NVM transistor 100 is negatively biased. The localized heat and negative bias causes an accelerated threshold voltage shift within p-channel NVM transistor 100.

[0031] In the present embodiment, p-channel NVM transistor 100 is programmed by applying a first positive supply voltage to the first end 110A of gate electrode 110, a second positive supply voltage to n-well 202, and a ground supply voltage to the second end 110B of gate electrode 110. The first positive

supply voltage is a core supply voltage V_{CC} , which is typically used to operate the core logic of a CMOS device, or an input/output supply voltage V_{DD} , which is typically used to operate the input/output logic of a CMOS device. For example, core supply voltage V_{CC} may have a nominal value of 1 Volt. The second positive supply voltage can be an input/output supply voltage V_{DD} , which is typically used to operate the input/output logic of a CMOS device. For example, input/output supply voltage V_{DD} may have a nominal value of 2.5 or 3.3 Volts. Note that both supply voltages V_{CC} and V_{DD} are normal operating voltages of the CMOS device.

[0032] Under the above-described conditions, a current of about 10 to 20 milli-Amperes passes through gate electrode 110, thereby heating the polysilicon and metal polysilicide layers of gate electrode 110. Also under these conditions, the second end 110B of gate electrode 110 is negatively biased with respect to n-well region 202. The heat generated within gate electrode 110 and the negative bias voltage on this gate electrode 110 results in a bias temperature degradation effect on p-channel NVM transistor 100. In accordance with one embodiment, the programming operation is performed for a duration of about 100 microseconds to 100 milliseconds. During this time, gate electrode 110 reaches a temperature in the range of about 400°C to 900°C. Under these conditions, the threshold voltage of transistor 100 shifts (i.e., becomes more negative), such that transistor 100 will not conduct current when a conventional read control voltage (e.g., 0 Volts) is applied to gate electrode 110. For example, the threshold voltage of p-channel transistor 100 may shift from a range of about -0.3 to -0.4 Volts, to a range of about -1.1 to -2.0 Volts. It is important to note that the programming operation permanently shifts the threshold voltage of p-channel

transistor 100, such that this transistor operates as a non-volatile memory cell.

[0033] To read the programmed/non-programmed state of p-channel NVM transistor 100, a ground supply voltage (0 Volts) is applied to both ends 110A and 110B of gate electrode 110 and to source/drain terminal 102. The V_{CC} supply voltage is applied to n-well region 202 and source/drain region 101. Current sense circuitry (not shown) is coupled to source/drain region 101. If NVM transistor 100 is programmed, the threshold voltage of this transistor is negative enough to prevent current from flowing between source/drain regions 101-102. Thus, the current sense circuitry fails to detect a significant read current when NVM transistor 100 is programmed. The current sense circuitry identifies the absence of read current as a first logic stage (e.g., a logic "1" value). Conversely, if NVM transistor 100 is not programmed, the threshold voltage of this transistor is less negative, thereby allowing a significant read current to flow between source/drain regions 101-102. Thus, the current sense circuitry detects a significant read current when NVM transistor 100 is not programmed. The current sense circuitry identifies the presence of a read current as a second logic state (e.g., a logic "0" value). In this manner, the current sense circuitry is able to identify the programmed/non-programmed state of p-channel NVM transistor 100. In the described embodiment, the duration of the read operation is comparable to the duration of a read operation of a conventional non-volatile memory cell, on the order of 1 microsecond.

[0034] P-channel NVM transistor 100 may be placed in an off (standby) state, wherein no current flows between source/drain regions 101-102, by applying the V_{CC} supply voltage to gate

electrode 110, source/drain region 101 and n-well region 202, and applying the ground supply voltage to source/drain region 102.

[0035] One advantage of p-channel NVM transistor 100 is that this transistor can be fabricated using a conventional CMOS process, without requiring any additional process steps or masks. In addition, NVM transistor 100 can be programmed without a high programming voltage (i.e., a voltage greater than the normal operating voltage of other transistors formed on the same substrate as transistor 100). NVM transistor 100 can advantageously be used in an application such as repairing bad circuitry (i.e., disabling faulty circuitry and enabling redundant circuitry). A plurality of NVM transistors identical to transistor 100 can be used to store an encryption key in a device for security applications. NVM transistor 100 has an advantage over conventional fuse-based NVM technology in an encryption key application. When using conventional fuse-based NVM technology, a visual inspection of the fuse-based elements may reveal the encryption key, because the physical characteristics of programmed and non-programmed fused-base NVM devices are different. However, the physical differences between the programmed and non-programmed NVM transistor 100 are subtle, thereby making it difficult to reverse engineer the contents of this NVM transistor 100.

[0036] Although the invention has been described in connection with several embodiments, it is understood that this invention is not limited to the embodiments disclosed, but is capable of various modifications, which would be apparent to one of ordinary skill in the art. For example, although a single p-channel NVM transistor has been described, it is understood that an array of these transistors can be created. Such an array could be used in any application where

a non-volatile memory fabricated using a standard CMOS process would be desirable, for instance in a programmable read-only memory (PROM) or a programmable logic device (PLD). As another example, other types of transistors, such as n-type transistors, may be used in accordance with the present invention. Thus, the present invention is only limited by the following claims.